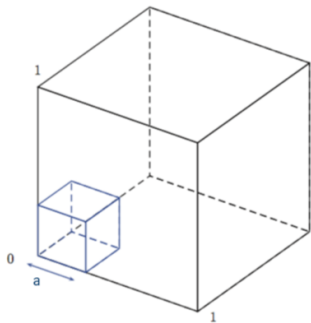**Introduction to Machine Learning**

Curse of Dimensionality
Curse of Dimensionality

Learning goals

- Understand that our intuition about geometry fails in high-dimensional spaces
- Understand the effects of the curse of dimensionality

CURSE OF DIMENSIONALITY

- The phenomenon of data becoming sparse in high-dimensional spaces is one effect of the **curse of dimensionality**.
- The **curse of dimensionality** refers to various phenomena that arise when analyzing data in high-dimensional spaces that do not occur in low-dimensional spaces.
- Our intuition about the geometry of a space is formed in two and three dimensions.
- We will see: This intuition is often misleading in high-dimensional spaces.
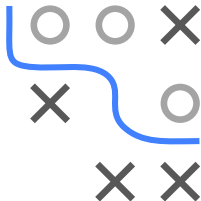
CURSE OF DIMENSIONALITY: EXAMPLE

To illustrate one of the problematic phenomena of data in high dimensional data, we look at an introductory example:

We are given 20 emails, 10 of them are spam and 10 are not. Our goal is to predict if a new incoming mail is spam or not.

For each email, we extract the following features:

- frequency of exclamation marks (in %)
- the length of the longest sequence of capital letters
- the frequency of certain words, e.g., "free" (in %)
- ...

... and we could extract many more features!

CURSE OF DIMENSIONALITY: EXAMPLE

Based on the frequency of exclamation marks, we train a very simple classifier (a decision stump with split point **x** = 0.25):

- We divide the input space into 2 equally sized regions.
- In the second region $[0.25, 0.5]$, 7 out of 10 are spam.
- Given that at least 0.25% of all letters are exclamation marks, an email is spam with a probability of $\frac{7}{10} = 0.7$.
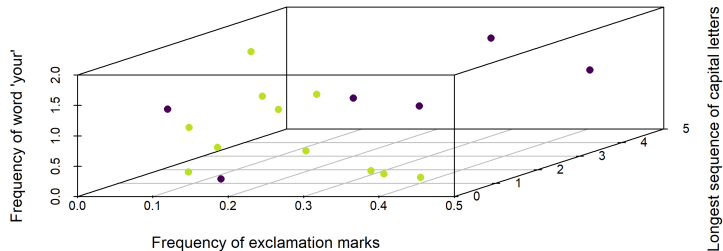
CURSE OF DIMENSIONALITY: EXAMPLE

Let us feed more information into our classifier. We include a feature
that contains the length of the longest sequence of capital letters.
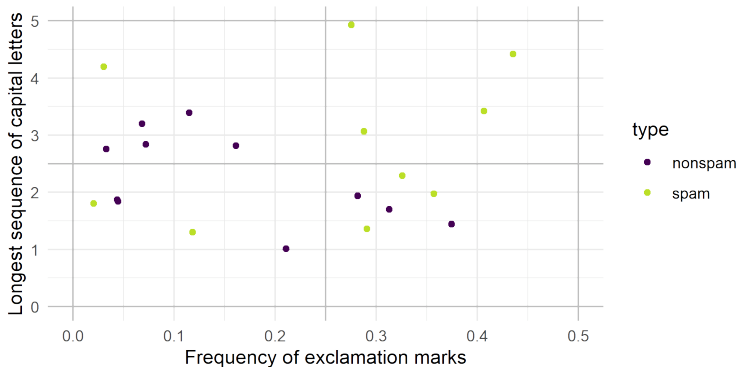
- In the 1D case we had 20 observations across 2 regions.
- The same number is now spread across 4 regions.

CURSE OF DIMENSIONALITY: EXAMPLE

Let us further increase the dimensionality to 3 by using the frequency of the word "your" in an email.

CURSE OF DIMENSIONALITY: EXAMPLE

- When adding a third dimension, the same number of observations is spread across 8 regions.
- In 4 dimensions the data points are spread across 16 cells, in 5 dimensions across 32 cells and so on ...
- As dimensionality increases, the data become **sparse**; some of the cells become empty.
- There might be too few data in each of the blocks to understand the distribution of the data and to model it.
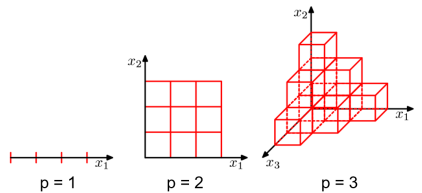
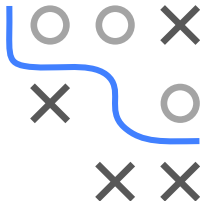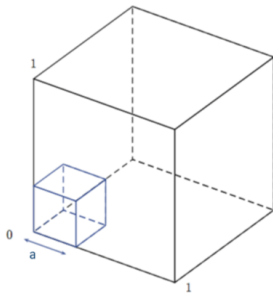

Bishop, Pattern Recognition and Machine Learning, 2006
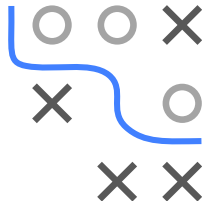
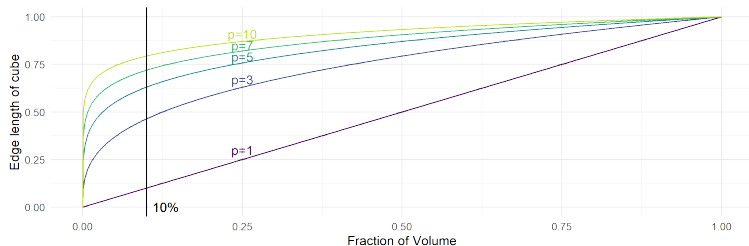**Geometry of High-Dimensional Spaces**
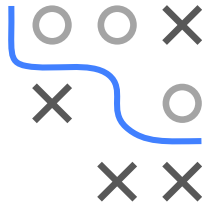
THE HIGH-DIMENSIONAL CUBE

- We embed a small cube with edge length *a* inside a unit cube.
- How long does the edge length *a* of this small hypercube have to be so that the hypercube covers 10%, 20%, ... of the volume of the unit cube (volume 1)?
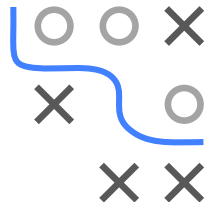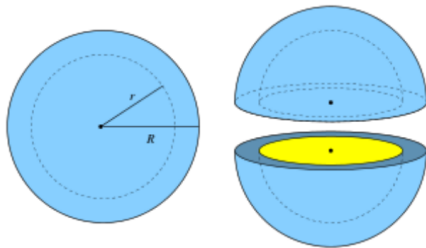
THE HIGH-DIMENSIONAL CUBE

$$a^p = \frac{1}{10} \Leftrightarrow a = \frac{1}{\sqrt[p]{10}}$$

- So: covering 10% of total volume in a cell requires cells with almost 50% of the entire range in 3 dimensions, 80% in 10 dimensions.

THE HIGH-DIMENSIONAL SPHERE

Another manifestation of the **curse of dimensionality** is that the majority of data points are close to the outer edges of the sample. Consider a hypersphere of radius 1. The fraction of volume that lies in the ϵ-"edge", $\epsilon := R - r$, of this hypersphere can be calculated by the formula
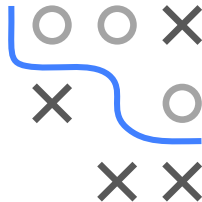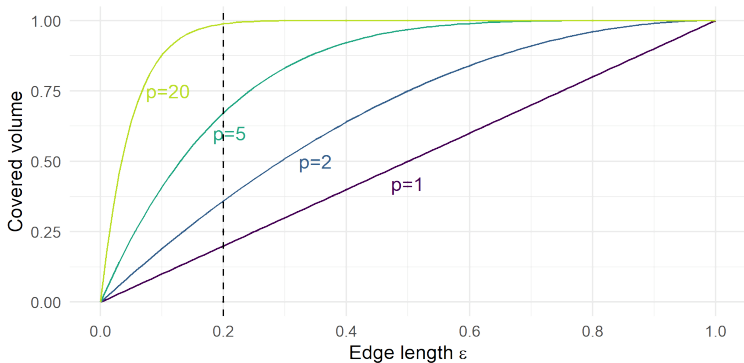
$$1 - \left(1 - \frac{\epsilon}{R}\right)^p.$$

If we peel a high-dimensional orange, there is almost nothing left.

THE HIGH-DIMENSIONAL SPHERE

Consider a 20-dimensional sphere. Nearly all of the volume lies in its outer shell of thickness 0.2:
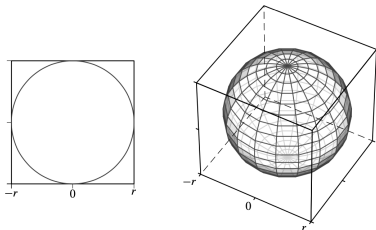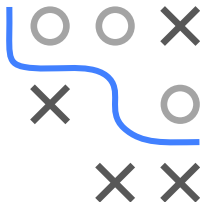
HYPHERSPHERE WITHIN HYPERCUBE

Consider a p-dimensional hypersphere of radius r and volume $S_p(r)$ inscribed in a p-dimensional hypercube with sides of length $2r$ and volume $C_p(r)$. Then it holds that

$$\lim_{p \to \infty} \frac{S_p(r)}{C_p(r)} = \lim_{p \to \infty} \frac{\left(\frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)}\right) r^p}{(2r)^p} = \lim_{p \to \infty} \frac{\pi^{\frac{p}{2}}}{2^p \Gamma(\frac{p}{2}+1)} = 0,$$

i.e., as the dimensionality increases, most of the volume of the hypercube can be found in its corners.
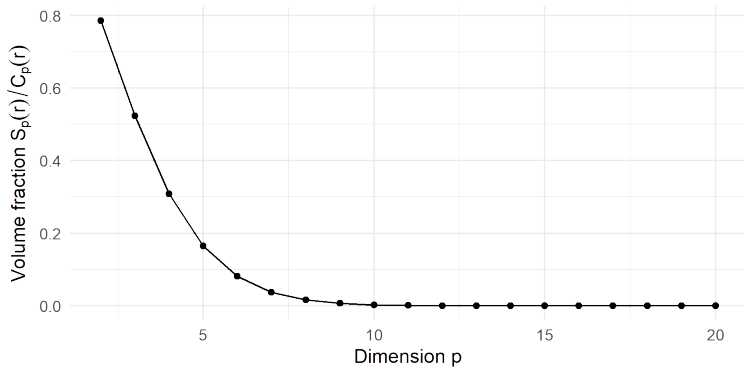

Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, 2014

HYPHERSPHERE WITHIN HYPERCUBE

Consider a 10-dimensional sphere inscribed in a 10-dimensional cube.
Nearly all of the volume lies in the corners of the cube:

Note: For $r > 0$, the volume fraction $\frac{S_p(r)}{C_p(r)}$ is independent of r.

UNIFORMLY DISTRIBUTED DATA

The consequences of the previous results for uniformly distributed data in the high-dimensional hypercube are:

- Most of the data points will lie on the boundary of the space.
- The points will be mainly scattered on the large number of corners of the hypercube, which themselves will become very long spikes.
- Hence the higher the dimensionality, the more similar the minimum and maximum distances between points will become.
- This degrades the effectiveness of most distance functions.
- Neighborhoods of points will not be local anymore.
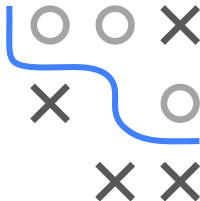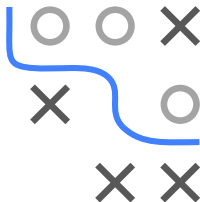
GAUSSIANS IN HIGH DIMENSIONS

A further manifestation of the **curse of dimensionality** appears if we consider a standard Gaussian $N_p(\mathbf{0}, \mathbf{I}_p)$ in p dimensions.

- After transforming from Cartesian to polar coordinates and integrating out the directional variables, we obtain an expression for the density $p(r)$ as a function of the radius r (i.e., the point's distance from the origin), s.t.

$$p(r) = \frac{S_p r^{p-1}}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{r^2}{2\sigma^2}\right),$$

where S_p is the surface area of the p-dimensional unit hypersphere.

- Thus $p(r)\delta r$ is the approximate probability mass inside a thin shell of thickness δr located at radius r.
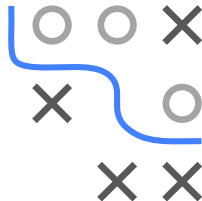
GAUSSIANS IN HIGH DIMENSIONS

- To verify this functional relationship empirically, we draw 10^4 points from the p-dimensional standard normal distribution and plot $p(r)$ over the histogram of the points' distances to the origin:
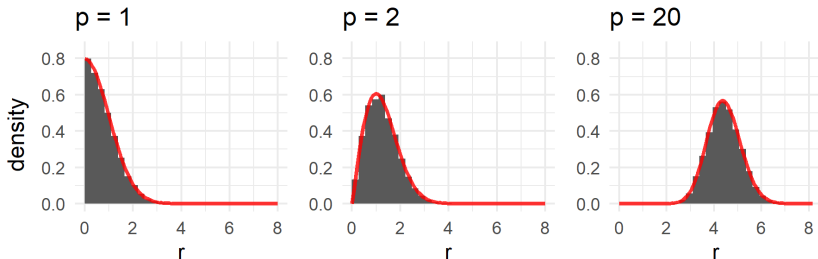
- We can see that for large *p* the probability mass of the Gaussian is concentrated in a fairly thin "shell" rather far away from the origin. This may seem counterintuitive, but:
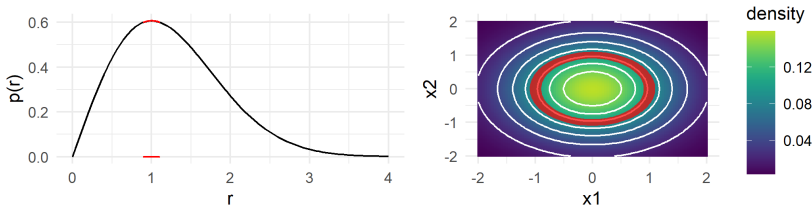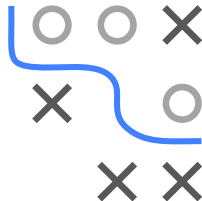
GAUSSIANS IN HIGH DIMENSIONS

- For the probability mass of a hyperspherical shell it follows that

$$\int_{r-\frac{\delta r}{2}}^{r+\frac{\delta r}{2}} p(\tilde{r})d\tilde{r} = \int_{r-\frac{\delta r}{2} \leq ||\mathbf{x}||_2 \leq r+\frac{\delta r}{2}} f_p(\tilde{\mathbf{x}})d\tilde{\mathbf{x}},$$

where $f_p(\mathbf{x})$ is the density of the p-dimensional standard normal distribution and $p(r)$ the associated radial density.

Example: 2D normal distribution

- While f_p becomes smaller with increasing r, the region of the integral -the hyperspherical shell- becomes bigger.

INTERMEDIATE REMARKS

However, we can find effective techniques applicable to
high-dimensional spaces if we exploit these properties of real data:

- Often the data is restricted to a manifold of a lower dimension.
  (Or at least the directions in the feature space over which
  significant changes in the target variables occur may be confined.)

- At least locally small changes in the input variables usually will
  result in small changes in the target variables.